УДК 519.237.8

Г. И. Турканов[1], *Е. В. Щепин*[1,2]

[1]Московский физико-технический институт (государственный университет)
[2]Математический институт имени В. А. Стеклова РАН

Классификатор Байеса для переменного количества признаков

Рассматривается подход ранжирования при помощи наивного байесовского классификатора для переменного количества признаков с применением теории фракталов, которая позволяет получить дополнительную информацию в классификатор-характеристику самоподобия. Для этого будет модифицирован наивный Байесовский классификатор и определен показатель Херста данных, который связан с традиционной фрактальной размерностью.

Ключевые слова: Байесовский классификатор, машинное обучение, ранжирование, показатель Херста, фрактальная размерность, предсказание, вычислительный эксперимент.

G. I. Turkanov[1], *E. V. Scepin*[1,2]

[1]Moscow Institute of Physics and Technology (State University)
[1,2]Steklov Mathematical Institute of Russian Academy of Sciences, Moscow, Russia

Bayes classifier for a variable number of features

The ranking approach using the Bayesian classifier for a variable number of features with the factual theory, which allows us to add more information to the classifier - characteristics of selfsimilarity. For this the Naive Bayes classifier is modified and defins Hurst data that is associated with traditional fractal dimension.

Key words: Bayesian classifier, machine learning, ranking, Hurst exponent, fractal dimension, prediction, computational experiment.

1. Введение

Так называемый *наивный классификатор Байеса* основан на предположении независимости рассматриваемых признаков [1]. А именно, оценка вероятности события по данной совокупности признаков, согласно Байесу, основана на произведении условных вероятностей этого события относительно рассматриваемых признаков.

И несмотря на то, что при практическом применении признаки, как правило, зависимы и байесовские оценки вероятностей могут сильно отличаться от реально наблюдаемых, этот простейший классификатор редко удается существенно превзойти [2].

Для практического применения байесовской оценки первоочередное значение имеет не ее точность, а ее ковариантность наблюдаемой вероятности, выражающаяся в том, что эта оценка тем больше, чем больше наблюдаемая вероятность события [3].

Целью исследований было увеличить точность классификатора при условии переменного количества признаков.

2. Байесовский классификатор

Существенным условием выполнения свойства ковариантности байесовской оценки является постоянство количества признаков, на основе которых она определяется.

Вероятностная модель для классификатора – это условная модель $p(C|x_1, x_2, ..., x_n)$ над независимой переменной класса C и признаками $x_1, x_2, ..., x_n$ по теореме Байеса [4]:

$$p(C|x_1, x_2, ..., x_n) = \frac{p(C)p(x_1, x_2, ..., x_n|C)}{p(x_1, x_2, ..., Fx_n)}.$$

В свою очередь знаменатель представляет собой масштабный множитель, зависящий только от признаков $x_1, x_2, ..., x_n$, числитель же эквивалентен совместной вероятности модели:

$$p(C)p(x_1, x_2, ..., x_n|C) = p(C)p(x_1|C)p(x_2|C, x_1)...p(x_n|C, x_1, x_2, ..., x_{n-1}).$$

Далее, в наивном байесовском классификаторе используется предположение о независимости признаков $x_1, x_2, ..., x_n$, то есть

$$R(x, C) \equiv p(C)p(x_1, x_2, ..., x_n|C) = p(C)\prod_{i=1}^{n} p(x_i|C).$$

В случае с переменным количеством признаков использование такого классификатора в явном виде приводит к потере его ковариантности, то есть объект с большим количеством признаков получит заведомо заниженную оценку.

Самое простое, что можно сделать в случае переменного количества признаков – это перейти от произведения к среднему геометрическому условных вероятностей. В случае постоянного количества признаков переход от произведения к среднему геометрическому никак не отражается на ковариантности классификатора, а в случае переменного количества, очевидно, ее повышает:

$$R^*_n(x, C) = \frac{1}{n}p(C)\prod_{i=1}^{n} p(x_i|C).$$

Для дальнейшего повышения ковариантности классификатора в случае переменного количества признаков предложено проанализировать характер зависимости байесовского произведения от количества множителей.

Если общее количество признаков велико (тысячи), тогда как для классификации каждого события применяется лишь небольшая их часть (десятки), то логично ожидать, что логарифм байесовского произведения асимптотически линейно растет с ростом количества сомножителей.

А именно, пусть $B(k, n)$ обозначает среднее значение логарифма байесовского произведения для k наблюдаемых событий с n признаками:

$$B(k, n) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n}p(C)\prod_{i=1}^{n} p(x_i^j|C) = \frac{1}{nk}p(C)\sum_{j=1}^{k}\prod_{i=1}^{n} p(x_i^j|C),$$

тогда на практике для зависимых признаков нередко можно наблюдать, что разность $B(k, n) - nB(k, 1)$ растет пропорционально некоторой (обычно нецелой) степени количества признаков:

$$B(k, n) - nB(k, 1) \sim cn^H. \quad (1)$$

Показатель H этой степени называется показателем Херста.

Далее будет показана модификация байесовского классификатора на основе следующего члена асимптотического разложения логарифма байесовского произведения.

3. Показатель Херста

Известно, что показатель Херста представляет собой меру персистентности — склонности процесса к трендам (в отличие от обычного броуновского движения). Значение $H > \frac{1}{2}$ означает, что направленная в определенную сторону динамика процесса в прошлом, вероятнее всего, повлечет продолжение движения в том же направлении. Если $H < \frac{1}{2}$, то прогнозируется, что процесс изменит направленность. $H = \frac{1}{2}$ означает неопределенность — броуновское движение [5].

Рассмотрим систему наблюдений $(x^n, y)_k$, где $\mathbf{x^n}$ – вектор признаков длины $n \in \{n_1, n_2, ..., n_N\}$, y – класс из $\{0, 1\}$, k – количество наблюдений. Предположим, что значения $\mathbf{x^n}$ так же принимают значения из $\{0, 1\}$.

Сперва перейдем к системе $(x_j, ctr_j)_{j=1..F}$, где F – общее количество признаков, а ctr_j:

$$ctr_j = \frac{1}{k_j + 1/ctr_0}(\sum_{i=1}^{k_j} y_i + 1),$$

где k – количество наблюдений, в которых встречается признак x_j, y_i – соответствующие этим наблюдениям классы.

Далее вычисляется среднее байесовских оценок для различного числа признаков n:

$$E = \frac{1}{|\{n_1, ..., n_N\}|} \sum_{n \in \{n_1, ..., n_N\}} R_n^*(x, C).$$

Затем рассчитывается стандартное отклонение:

$$\sigma_n = \sqrt{\frac{1}{|\{n_1, ..., n_N\}|} \sum_{n \in \{n_1, ..., n_N\}} (R_n^*(x, C) - E)^2}.$$

И окончательно в предположении (1) –

$$\ln \frac{r_n}{\sigma_n} \sim H \ln n.$$

Откуда угол наклона прямой, построенной как аппроксимация последовательности $\frac{r_n}{\sigma_n}$:

$$\frac{r_n}{\sigma_n} \sim \left(\frac{n}{c}\right)^H,$$

где

$$r_n = \max_n \frac{1}{k_n} \sum_{t=1}^{k_n} ctr_t^n - \min_n \frac{1}{k_n} \sum_{t=1}^{k_n} ctr_t^n$$

Угол наклона H прямой:

$$H \log \left(\frac{n}{\mathrm{c}}\right) - \log \frac{r_n}{\sigma_n} = 0$$

– искомый показатель Херста.

Дополнительная информация, которую несет показатель H как коэффициент самоподобия, далее была применена к байесовскому классификатору:

$$R^{**} = \frac{1}{n^H} p(C) \prod_{i=1}^{n} \frac{p(x_i|C)}{E}.$$

4. Экспериментальные результаты

В качестве экспериментальной базы использованы данные поисковой системы Yandex. Обучающая выборка была собрана за период с 27.10.2015 по 16.11.2015 и состоит из 24 099 318 пар фраза–баннер с общим числом 440 205 425 показов и 6 685 997 кликов.

Тестовый набор данных был собран с 17.11.2015 по 23.11.2015 и включает в себя 1 139 066 пар фраза–баннер с общим числом 7 182 355 показов и 40 877 кликов.

Метрикой качества выступает количество потерянных кликов по рекламным баннерам в зависимости от порога фильтруемых показов. На графике изображена разница между потерянными кликами, зеленый – наивный байесовский классификатор, синий – модифицированный. Отрицательные значения соответствуют меньшему значению потерянных кликов при одинаковом значении фильтруемых показов.

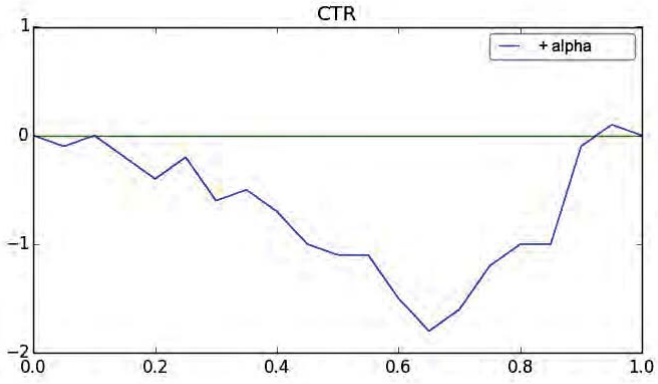

Рис. 1. Разница в фильтрованных кликах по наивному байесовскому классификатору и модифицированному

В качестве альтернативной метрики была использована ROC-кривая. Для наивного байесовского классификатора значение площади под кривой составило 0.721377, для модифицированного метода – 0.760481. В качестве бинарной классификации выступало наличие или отсутствие клика в паре фраза–баннер.

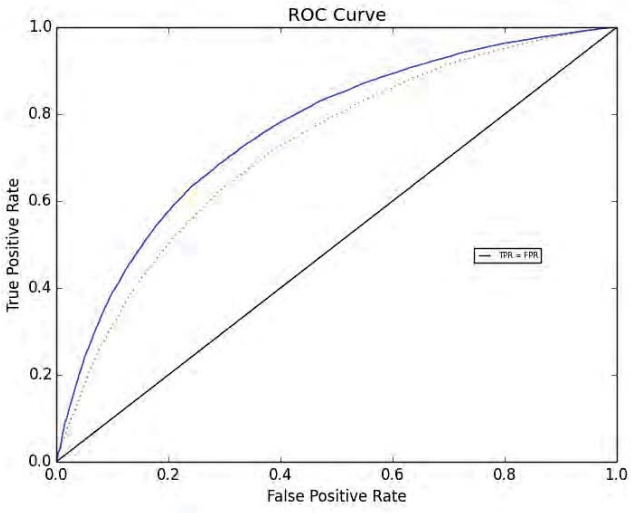

Рис. 2. ROC-кривые двух классификаторов, синий – модифицированный метод

5. Интерпретация результатов

Экспериментальные результаты подтвердили лежащие в основе исследования предположения о положительном вкладе дополнительной информации в предсказание в виде фрактальной размерности.

Показатель H составил $0,68$, что говорит о периодичной зависимости в данных. Улучшение в предсказании наблюдается на всем промежутке фильтруемых показов рекламных баннеров, относительное улучшение до 6% кликов.

Для подтверждения предположения о связи между показателем H и зависимостью используемых признаков был произведен еще один эксперимент. В исходные данные было добавлено искажение в виде дублирования 20% признаков, что привело к увеличению показателя H до $0,81$ и ухудшения классификатора по сравнению с более независимыми признаками.

Литература

1. *Russell S., Norvig P.* Artificial Intelligence: A Modern Approach (2nd ed.). New York: Prentice Hall, 2003.
2. *Graepel T., Candela J., Borchert T., Herbrich R.* Web-Scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine // Proceedings of 27th International Conference on Machine Learning. 2010. P. 13–20.
3. *Turkanov G.* Modified Naive Bayes with Hurst exponent as quantitative measure of data mutual dependence // Proceedings of Yandex School of Data Analysis Conference. Machine Learning: Prospects and Applications. 2015.
4. *Bayes T.* An essay, towards solving a problem in the doctrine of chances // Philos Trans R Soc Lond. 1763. V. 53. P. 370–418.
5. *Feder J.* Fractals. New York: Plenum Press, 1988.

References

1. *Russell S., Norvig P.* Artificial Intelligence: A Modern Approach (2nd ed.). New York: Prentice Hall, 2003.
2. *Graepel T., Candela J., Borchert T., Herbrich R.* Web-Scale Bayesian click-through rate prediction for sponsored search advertising in Microsoft's Bing search engine. Proceedings of 27th International Conference on Machine Learning. 2010. P. 13–20.
3. *Turkanov G.* Modified Naive Bayes with Hurst exponent as quantitative measure of data mutual dependence. Proceedings of Yandex School of Data Analysis Conference. Machine Learning: Prospects and Applications. 2015.
4. *Bayes T.* An essay, towards solving a problem in the doctrine of chances. Philos Trans R Soc Lond. 1763. V. 53. P. 370–418.
5. *Feder J.* Fractals. New York: Plenum Press, 1988.

Поступила в редакцию 23.09.2016